DOCUMENT RESUME

ED 428 101 TM 029 502

AUTHOR Glas, Cees A. W.; Vos, Hans J.
TITLE Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory. Research Report 98-15.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE 1998-00-00
NOTE 43p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Bayesian Statistics; *Item Response Theory; *Mastery Tests; Simulation; Tables (Data); Test Items
IDENTIFIERS Decision Theory; *Rasch Model; Sequential Data Analysis; Testlets; Three Parameter Model; Two Parameter Model

ABSTRACT

A version of sequential mastery testing is studied in which response behavior is modeled by an item response theory (IRT) model. First, a general theoretical framework is sketched that is based on a combination of Bayesian sequential decision theory and item response theory. A discussion follows on how IRT based sequential mastery testing can be generalized to adaptive item and testlet selection rules; i.e., to a situation in which the choice of the next item or testlet to be administered is optimized using the information from previous responses. The performance of IRT based sequential and adaptive sequential mastery testing is studied in a number of simulations using the Rasch model. The possibilities and difficulties of application of the approach in the framework of the two-parameter logistic and three-parameter logistic models is also discussed. (Contains 19 tables and 16 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 428 101

TM

Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory

Research Report 98-15

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Nelissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Cees A.W. Glas
Hans J. Vos

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

TM029502

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

2

Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory

Cees A.W. Glas

Hans J. Vos

Department of Educational Measurement and Data Analysis
Faculty of Educational Science and Technology
University of Twente, Enschede, the Netherlands
E-mail C.A.W.Glas@edte.utwente.nl

Abstract

In this paper, a version of sequential mastery testing is studied where response behavior is modeled by an item response theory (IRT) model. Firstly, a general theoretical framework will be sketched that is based on a combination of Bayesian sequential decision theory and item response theory. Then it will be pointed out how IRT based sequential mastery testing can be generalized to adaptive item and testlet selection rules, that is, to a situation where the choice of the next item or testlet to be administered is optimized using the information from previous responses. The performance of IRT based sequential and adaptive sequential mastery testing will be studied in a number of simulations using the Rasch model. Finally, the possibilities and difficulties of application of the approach in the framework of the 2-PL and the 3-PL model will be discussed.

Key words: adaptive testing, Bayesian sequential decision theory, mastery testing, item response theory, testlets.

Introduction

In mastery testing the problem is to decide on either mastery or non-mastery, given an examinee's observed response pattern. Well-known examples of mastery testing include pass/fail decisions, licensure, and certification. The mastery test can have both a fixed-length and variable-length format. In the fixed-length mastery test, the performance on a fixed number of items is used for deciding on either mastery or non-mastery. Over the last few decades, the fixed-length mastery problem has been studied extensively by many researchers (e.g., De Gruijter & Hambleton, 1984; van der Linden, 1990). Most of these authors derived, analytically or numerically, optimal rules by applying (empirical) Bayesian decision theory (e.g. DeGroot, 1970; Lehmann, 1986) to this problem.

In the variable-length format, in addition to the actions declaring mastery or non-mastery, also the action continuing and administering another item is available (e.g., Kingsbury & Weiss, 1983; Lewis & Sheehan, 1990; Sheehan & Lewis, 1992; Spray & Reckase, 1996). The main advantage of variable-length mastery tests as compared to fixed-length mastery tests is that they offer the possibility to provide shorter tests for those examinees who have clearly attained a certain level of mastery (or clearly non-mastery) and longer tests for whom the mastery decision is not as clear-cut (Lewis & Sheehan, 1990). For instance, Lewis and Sheehan (1990) showed in a simulation study that average test lengths could be reduced by a half without sacrificing classification accuracy.

Two main types of variable-length or multistage mastery tests can be distinguished. First, the next item to be administered can be selected random. In this case, the stopping rule (i.e., termination criterion) is adaptive but the item selection procedure is not adaptive. This type of variable-length mastery problem is also known as a sequential mastery problem, and, in the sequel, it will be referred to as SMT. Examinees with a low and high level of ability are classified as non-master and master, respectively, whereas those with an intermediate level of ability are presented another item to be randomly selected. In case the termination criterion is determined using Bayesian sequential decision theory (e.g., DeGroot, 1970; Lehmann, 1986), and a computer is used for selecting and scoring the next random item, Lewis and Sheehan (1990) denote this type of sequential mastery testing as computerized mastery testing. Costs of administering one random item can explicitly be taken into account within the framework of Bayesian sequential decision theory.

In the second main type of variable-length mastery testing not only the stopping rule, but also the item selection mechanism is adaptive . The examinee's ability level is estimated

after each response, and the item to be administered next to each examinee is neither too easy nor too difficult for that examinee. In other words, able examinees can avoid doing too many easy items and less able examinees can avoid being exposed to too many difficult items. Kingsbury and Weiss (1983) denote this type of variable-length mastery testing as adaptive mastery testing and in the sequel, this will be referred to as ASMT (adaptive sequential mastery testing). In ASMT it is assumed that items have unequal difficulty implying that the probability to answer an item correctly is not equal for all items in the pool. It should be noted that, although items are also allowed to have unequal difficulty in sequential mastery testing, the next item is randomly selected in this problem.

Sequential Mastery Testing

In this section, a general theoretical framework for SMT will be presented that is based on a combination of Bayesian sequential decision theory and item response theory (IRT). This framework is an extension of the approach by Lewis and Sheehan (1990). Consider a situation where one must decide whether or not a person has such an ability level that he or she can be considered a master. So let θ_c be some cut-off point on a latent continuum, persons with ability θ below this cut-off point are non-masters, persons with ability θ above this cut-off point are masters. To make the decision, a number of testlets consisting of one or more items are administered. Suppose that the procedure consists of S stages labeled $s = 1, ..., S$ and at each stage one of the testlets can be given. Then, at stage s, $s < S$, three decisions can be made:

$$d = \begin{cases} m & \text{the respondent is judged a master, sampling stops,} \\ n & \text{the respondent is judged a non-master, sampling stops,} \\ c & \text{sampling is continued.} \end{cases} \tag{1}$$

So in the first two cases administering testlets is terminated, while in the third case, a new testlet is given. The loss associated with the first two decisions is

$$L(m, \theta) = max\{sC, sC + A(\theta - \theta_c)\} \tag{2}$$

with $A < 0$ and

$$L(n, \theta) = max\{sC, sC + B(\theta - \theta_c)\}. \tag{3}$$

with $B > 0$; C is the cost of delivering one testlet, sC is the cost of delivering s testlets.

The decision will be based on the response pattern of the respondent. Let $\mathbf{x}_s$ be the response to the s-th testlet. Further, define the response patterns $\mathbf{y}_s = (\mathbf{x}_1, \ldots, \mathbf{x}_s)$, for $s = 1, \ldots, S$. At stage s, the decision d, that is, the decision whether the respondent is a master or a non-master, or whether another testlet will be administered, will be based on the expected losses of the three possible decisions given the observed response pattern $\mathbf{y}_s$. The expected losses of the first two decisions given a response pattern $\mathbf{y}_s$ are computed as

$$E(L(m,\theta) \mid \mathbf{y}_s) = sC + A \int_{-\infty}^{\theta_c} (\theta - \theta_c) p(\theta \mid \mathbf{y}_s) d\theta \tag{4}$$

and

$$E(L(n,\theta) \mid \mathbf{y}_s) = sC + B \int_{\theta_c}^{\infty} (\theta - \theta_c) p(\theta \mid \mathbf{y}_s) d\theta, \tag{5}$$

where $p(\theta \mid \mathbf{y}_s)$ is the posterior density of θ given $\mathbf{y}_s$. The expected loss of the third possible decision is computed as the expected risk of decisions taken in the follow-up testlets. Let $\{\mathbf{x}_{s+1}\}$ be the set of all possible response patterns on testlet $s+1$. Then, for $s = 1, \ldots, S-1$, the expected risk of continuing testing is defined as

$$E(R(\mathbf{x}_{s+1}, \mathbf{y}_s) \mid \mathbf{y}_s) = \sum_{\{\mathbf{x}_{s+1}\}} p(\mathbf{x}_{s+1} \mid \mathbf{y}_s) R(\mathbf{x}_{s+1}, \mathbf{y}_s) \tag{6}$$

with the so-called posterior predictive distribution $p(\mathbf{x}_{s+1} \mid \mathbf{y}_s)$ given by

$$p(\mathbf{x}_{s+1} \mid \mathbf{y}_s) = \int p(\mathbf{x}_{s+1} \mid \theta) p(\theta \mid \mathbf{y}_s) d\theta \tag{7}$$

and risk defined as

$$R(\mathbf{x}_{s+1}, \mathbf{y}_s) = min\{E(L(m,\theta) \mid \mathbf{y}_{s+1}), E(L(n,\theta) \mid \mathbf{y}_{s+1}), E(R(\mathbf{x}_{s+1}, \mathbf{y}_s) \mid \mathbf{y}_s)\}. \tag{8}$$

The risk associated with the last testlet is defined as

$$R(\mathbf{x}_S, \mathbf{y}_{S-1}) = min\{E(L(m,\theta) \mid \mathbf{y}_S), E(L(n,\theta) \mid \mathbf{y}_S)\}. \tag{9}$$

Notice that the definitions (6) through (9) imply a recursive definition of the expected risk of

continuation. Since evaluation of $E(R(\mathbf{x}_{s+1},\mathbf{y}_s) \mid \mathbf{y}_s)$ entails a summation over the set of all possible response patterns $\{\mathbf{x}_{s+1}\}$, exact computation of this expected risk generally presents a major problem. One of the approaches to this problem is approximating (6) using Monte Carlo simulation techniques, that is, simulating a large number of draws from $p(\mathbf{x}_{s+1} \mid \mathbf{y}_s)$ to compute the mean of $R(\mathbf{x}_{s+1},\mathbf{y}_s)$ over these draws. This approach is beyond the scope of the present paper and will be treated later. However, in the case that the IRT model for $\mathbf{x}_{s+1}$, say $p(\mathbf{x}_{s+1} \mid \theta)$, defines an exponential family, the problem of the large number of possible response patterns is solved by the existence of minimal sufficient statistics. An example will be given in the next section.

The Rasch model

In the Rasch model the probability of a response pattern $\mathbf{x}$ on a test of K items is given by

$$\begin{aligned} p(\mathbf{x} \mid \theta,\boldsymbol{\beta}) &= \prod_{i=1}^{K} \frac{\exp(x_i(\theta-\beta_i))}{1+\exp(\theta-\beta_i)} \\ &= \exp(t\theta)\exp(-\mathbf{x}'\boldsymbol{\beta})P_0(\theta), \end{aligned} \tag{10}$$

where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K)$ is a vector of item parameters, $t = \sum_i x_i$ is the sum score and

$$P_0(\theta) = \prod_{i=1}^{K}(1+\exp(\theta-\beta_i))^{-1}. \tag{11}$$

Notice that t is the minimal sufficient statistic for θ. Further, it is easily verified that $P_0(\theta)$ is the probability, given θ, of a response pattern with all item responses equal to zero. The probability of observing t given θ is given by

$$\begin{aligned} p(t \mid \theta) &= \sum_{\{\mathbf{x}|t\}} \exp(t\theta-\mathbf{x}'\boldsymbol{\beta})P_0(\theta) \\ &= \gamma_t(\boldsymbol{\beta})\exp(t\theta)P_0(\theta) \end{aligned}$$

with γ_t an elementary symmetric function defined by $\gamma_t = \sum_{\{\mathbf{x}|t\}}\exp(-\mathbf{x}'\boldsymbol{\beta})$ where $\{\mathbf{x} \mid t\}$

stands for the set of all possible response patterns resulting in a sum score t. An important feature is that the posterior distributions of θ given $\mathbf{x}$ and t are the same, that is

$$\begin{aligned} p(\theta \mid \mathbf{x}) &= \frac{\exp(t\theta - \mathbf{x}'\boldsymbol{\beta})P_0(\theta)g(\theta)d\theta}{\int \exp(t\theta - \mathbf{x}'\boldsymbol{\beta})P_0(\theta)g(\theta)d\theta} \\ &= \frac{\exp(t\theta)P_0(\theta)g(\theta)d\theta}{\int \exp(t\theta)P_0(\theta)g(\theta)d\theta} \\ &= \frac{\gamma_t(\boldsymbol{\beta})\exp(t\theta)P_0(\theta)g(\theta)d\theta}{\int \gamma_t(\boldsymbol{\beta})\exp(t\theta)P_0(\theta)g(\theta)d\theta} = p(\theta \mid t). \end{aligned}$$

At this point, an assumption will be introduced that may not be completely realistic. It will be assumed that local independence simultaneously holds within and over testlets, than is, all item responses are independent given θ. So at this point, no special requirements are made to model a possible dependence structure of testlet responses, this point will be returned to later. Then analogously to the sequential testing procedure described above, the posterior distribution of θ given a response pattern $\mathbf{y}_s$, $p(\theta \mid \mathbf{y}_s)$, is equivalent to the posterior of θ given a score pattern $\mathbf{t}_s$, $\mathbf{t}_s = (t_1, ..., t_s)$, in fact, it is equivalent to the posterior of θ given a score r_s, $r_s = \sum_{p=1}^{s} t_p$. Let $p(\theta \mid r_s)$ stand for the latter density. As a result, the expected losses (4), (5) and the expected risk (6) can be written as $E(L(m,\theta) \mid r_{s+1})$, $E(L(n,\theta) \mid r_{s+1})$ and $E(R(t_{s+1},r_s) \mid r_s)$. More specifically, the last loss is given by

$$E(R(t_{s+1}, r_s) \mid r_s) = \sum_{t_{s+1}=0}^{K_{s+1}} p(t_{s+1} \mid r_s) R(t_{s+1}, r_s) \tag{12}$$

and (7) specializes to

$$\begin{aligned} p(t_{s+1} \mid r_s) &= \int p(t_{s+1} \mid \theta) p(\theta \mid r_s) d\theta \\ &= \int \gamma_{t_{s+1}}(\boldsymbol{\beta}_{s+1}) \exp(t_{s+1}\theta) P_{0(s+1)}(\theta)\, p(\theta \mid r_s) d\theta, \end{aligned} \tag{13}$$

where $\boldsymbol{\beta}_{s+1}$ is a vector of the item parameters of testlet $s+1$ and $P_{0(s+1)}(\theta)$ is equal to (11) evaluated using $\boldsymbol{\beta}_{s+1}$, that is, $P_{0(s+1)}(\theta)$ is equal to the probability of a zero response pattern on testlet $s+1$, given θ. Since elementary functions can be very quickly computed up to any degree of precision (Verhelst, Glas and van der Sluis, 1984), the risk functions can be explicitly

computed.

Adaptive Sequential Mastery Testing

One of the topics addressed in this study is how the sequential testing procedure can be optimized in case a large testlet bank is available. So the question is which testlets must be administered next upon observing $\mathbf{y}_s$. Three approaches will be considered. The first two are directly taken from the framework of non-Bayesian adaptive mastery testing (see, for instance, Kingsbury and Weiss, 1983, Weiss and Kingsbury, 1984). Both are based on the maximum information criterium; the first approach entails choosing items or testlets with maximum information at θ_c, the second one with maximum information at $\hat{\theta}_s$, which is an estimate of θ at stage s. The third approach relates to a distinct difference between the non-Bayesian and Bayesian approach. In the former approach, one is interested in a point estimate of θ or in the question whether θ is below or above some cut-off point. In the latter approach, however, one is primarily interested in minimizing possible losses due to misclassifications and the costs of testing. This can be directly translated into a selection criterium for the next testlet. In a Bayesian framework for traditional computer adaptive testing, one might be interested in the posterior expectation of θ. One of the selection criteria suited for optimizing testlet administration is choosing the testlet with the minimal expected posterior variance. So if $\mathbf{y}_s$ is the observed response pattern, and $\{\mathbf{x}_{s+1}\}$ is the set of all possible response patterns on the next testlet, one may select the testlet where

$$\sum_{\{\mathbf{x}_{s+1}\}} var(\theta \mid \mathbf{y}_s, \mathbf{x}_{s+1}) p(\mathbf{x}_{s+1} \mid \mathbf{y}_s)$$

is minimal (see, for instance, van der Linden, 1998). In a sequential mastery testing framework, however, one is interested in minimizing possible losses, so a natural criterium for selection of the next testlet is

$$\sum_{\{\mathbf{x}_{s+1}\}} var(L(m,\theta) - L(n,\theta) \mid \mathbf{y}_s, \mathbf{x}_{s+1}) p(\mathbf{x}_{s+1} \mid \mathbf{y}_s), \tag{14}$$

that is, a testlet is chosen such that the expected reduction in the variance of the difference between the losses of the mastery and non-mastery decision is maximal. In other words, this criterium focuses on the posterior variance of θ given a response pattern $(\mathbf{y}_s, \mathbf{x}_{s+1})$, and the criterium entails that the sum over all possible follow-up response patterns $\mathbf{x}_{s+1}$ of this posterior

BEST COPY AVAILABLE

variance weighted by its posterior predictive probability $p(\mathbf{x}_{s+1} \mid \mathbf{y}_s)$ is minimal. In the case of the Rasch model, (14) is relatively easy to compute, because the response patterns $\mathbf{x}_{s+1}$ and $\mathbf{y}_s$ can be interchanged with the scores t_{s+1} and r_s. For the 2- and 3-PL, a simulation procedure similar to the procedure of the previous section can be adopted, but this is beyond the scope of the present paper.

Classification Precision and Adaptive Item Selection

Above, SMT was characterized by an adaptive stopping rule and ASMT by an adaptive stopping rule and an adaptive item selection method.

Before proceeding with presenting the results of a number of simulation studies of the performance of SMT and ASMT using the Bayesian decision theoretic-approach described above, a small study of adaptive mastery testing without an adaptive stopping rule but with an adaptive item selection method will be presented. One of the essential differences between this and the above framework will be the absence of a loss function involving the distance of a person's ability to the cut-off point and the cost of testing. The reason for this digression is to present some benchmark of how much adaptive item selection might improve decision accuracy in a context without an adaptive stopping rule, that is, in a context that is more like the usual practice of computer adaptive testing (CAT).

Consider a computer adaptive testing situation where items are selected using the maximum information criterium. In the Rasch model, item information is maximal if the item parameter equals the person parameter, that is, if the probability of a correct response given θ and β is equal to 0.50. In the sequel this probability will be called a local p-value. Suppose both a person's ability parameter and all item parameters are known. Suppose further that the item bank can support giving the optimal item time and again, so that an optimal test with all response probabilities equal to 0.50 could be given. In the second column of Table 1, the standard errors of θ for such a test are given for test lengths ranging from 10 to 100 items. These standard errors are computed as the square root of the inverse test information. In the third and fourth column, standard errors of θ are given for sub-optimal tests with all items response probabilities equal to 0.25 and 0.10, respectively.

Insert Table 1 about here

In Table 1 it can be seen that the difference between the standard error of an optimal test and

a sub-optimal test decreases with test length: for a test of 10 items the gain in precision from a test with items with local p-values of 0.25 to a test with local p-values of 0.50 is equal to the difference between 0.7303 and 0.6325, which is 0.0978. In the case of 100 items, this gain decreases to 0.0309. This, of course, is only an artificial example, in practice, adaptive testing is based on estimates of item and person parameters, and if the sub-optimal test is a paper-and-pencil test, there will be variation in the respondents' ability values and, as a consequence, in the local and overall p-values of the items. Still, the example shows that the expected gain in precision from adaptive testing must not be overestimated.

In Tables 2 to 4, a comparison is made of the gain in correct non-mastery classifications as a function of the cut-off point θ_c. Consider Table 2, where, for a test of 10 items, a comparison is made of the gain in proportion of correct decisions when moving from a sub-optimal test with a local p-value at $\theta = 0$ equal to 0.25 to an optimal test with local p-value at $\theta = 0$ of 0.50.

Insert Table 2 to 4 about here

In the first column, values of the cut-off point θ_c ranging from 0.00 to 1.90 are given. In the second and third column, the proportions of persons with an estimated ability above the cut-off point are given, for a test with local p-values of 0.50 and 0.25, respectively. So the entries in these two columns are the proportion of persons with true ability equal to zero that are incorrectly judged as masters. These proportions are computed using the standard errors of the first row of Table 1, and the assumption that $\hat{\theta}$ has a normal distribution. In the last column the gain in the proportion of correct responses is computed as the difference of the entries in the third and second column. Notice that the gain is single peaked, it is equal to zero if $\theta_c = 0$ and it goes to zero if θ_c goes to infinity. At $\theta_c = 0.7$ an optimum gain of 0.0347 is attained. Tables 3 and 4 contain analogous information for test lengths of $K = 20$ and $K = 40$. Notice that the magnitude of the optimum does not change much, only the position of the optimum moves closer to zero. As in the previous example, it must be stressed that this is a highly artificial example, but it cannot be expected that the gain will be much higher in a real life situation with estimates of item and person parameters and sub-optimal item banks.

Insert Table 5 about here

Finally, to obtain some flavor of the influence of the estimates of ability on the proportion of correct decisions the following simulation studies, reported in Table 5, were

BEST COPY AVAILABLE

carried out. The table consists of two panels, the first one pertains to studies every simulee had the same ability $\theta = 0.50$, the second panel pertains to studies where, for every simulee, a value θ was drawn from a standard normal distribution. The cut-off point in all studies was $\theta_c = 1.00$. The studies focussed on three variables: test length, ability estimation method and item selection method. Test lengths were chosen equal to 10, 20 and 40. In Table 5, the columns pertaining to these conditions are labeled "K=10", "K=20" and "K=40". Ability was either estimated with weighted maximum likelihood (Warm, 1989) or expected a-posteriori ability (Bock and Mislevy, 1982). The EAP estimate is computed using a standard normal prior. The columns pertaining to these conditions are labeled "WML" and EAP", respectively. Four item selection methods were studied. In the first one, for every simulee item parameters were randomly drawn from the standard normal distribution. In the second method, all item parameters were equal to θ_c. In the third method, all item parameters were equal to the true ability parameter of the testee. Finally, for the fourth method, the item parameter was equal to the current ability estimate θ_s. For all studies, the cut-off score θ_c was equal to 1.00. The rows of Table 5 pertaining to these four selection methods are labeled "random", "$\beta = \theta_c$", "$\beta = \theta$", and "$\beta = \theta_s$", respectively. Of course, the three last conditions are highly artificial, because it is assumed that the most informative item at θ_c, the true theta θ, and the running estimate θ_s is always available. Further, the true ability θ is unknown. Still, the simulations can be interesting reference material for the evaluation of the results on SMT and ASMT that will be given below. Table 5 contains the proportions correct decisions in 5000 replications for each combination of test length, ability estimation method and item selection method. The complete study was replicated several times, the standard errors of the proportions in the table are about 0.01. It can be seen that, for the studies with fixed θ, both item selection at the true θ and at θ_c produced the largest proportion of correct decisions. Selection at the running estimate of θ did not systematically produce results better than random item selection. In the second panel of Table 5, it can be seen that randomly drawing θ generally produced less favorable results. Further, the positive association between test length K and proportion of correct classifications vanished. This is due to the fact that random drawing of θ results in a lot of values far away from θ_c, where the proper classification is obvious after selection of only a few items and adding more items contributes little to classification precision. This leads to the expectation that in these cases, item and testlet administration will be quickly terminated in a sequential Bayesian mastery testing framework. In the following section, it will be investigated whether these expectations are justified.

Performance of Sequential and Adaptive Sequential Mastery Testing

Design of the study

In this section, the relation between various selection methods on one hand and the proportion of correct decisions, the proportions of testlets given and the mean loss on the other hand will be studied with a number of simulation studies. The main research questions will be whether, and under which circumstances, sequential testing improves upon a fixed test, and whether, and under which circumstances, adaptive sequential testing improves upon sequential testing. The design of the studies will be explained using the results of the first study, reported in Table 6.

Insert Tables 6 and 7 about here

The study concerns ten items and the cut-off point θ_c is equal to one. The nine bottom lines of the table represent nine simulation studies of 1000 replications each. For every replication a true θ was drawn from a standard normal distribution. In the first simulation study, every simulee was presented a fixed test of 10 items. For every simulee the item parameters were drawn from a standard normal distribution. Also the prior distribution of ability was standard normal. The next two sets of four conditions were two sequential mastery testing procedures, one with two testlets of five items and one with 10 testlets of one item. For these sequential mastery testing procedures, the parameters of the loss functions (2) and (3) were equal to $A = -1.00$, $B = 1.00$ and $C = 0.01k_t$, where k_t stands for the number of items in a testlet. The motivation for this choice of C is keeping the total cost of administering 10 items constant. The numbers of testlets and the numbers of items within testlets are summarized in the first two columns of Table 6. In the next row, the selection method is specified further. The two rows labeled "sequential" stand for a SMT condition were the item parameters of the first testlet were all equal to zero and the item parameters of all other testlets were randomly drawn from a standard normal distribution.

The conditions labeled "max info", min variance" and "cutting point" entail ASMT procedures. Also in these conditions the first testlet has all item parameters equal to zero. The reason for starting both the SMT and ASMT procedures with testlets with similar item parameters was to create comparable conditions in the initial phase of the procedures. The following testlets were chosen from a bank of 50 testlets that was generated as follows. First $50k_t$ item parameters were drawn from the standard normal distribution. Then, these $50k_t$

item parameters were ordered in magnitude from low to high. The first k_t items comprised the first testlet in the bank, the second k_t items comprised the second testlet, etcetera. In this way, 50 testlets were created that were homogeneous in difficulty and attained their maximum information at distinct points of the latent ability scale. In the "max info" condition, at stage s, $s = 1, ..., S - 1$, an expected a posteriori estimate of ability was computed and the expected risk of a "continue sampling" decision was computed using the $S - s$ testlets with highest information at this estimate. If a "continue sampling" decision was made, the next testlet administered was the most informative testlet of the $S - s$ testlets initially selected. The procedure in the "min variance" condition was roughly similar, only here the minimum variance criterium defined by (14) was used. Finally, in the "cutting point" condition, testlets were selected from the testlet bank described above that were most informative at the cutting point θ_c. The last three columns of Table 6 give the proportion of correct decisions, the proportion of testlets given and the mean loss over 1000 replications for each of the nine conditions, where the loss in every replication was computed using (2) or (3) evaluated at the true value of θ, with s as the number of testlets actually given.

Results

The study described in the previous section was carried out for three total test lengths, $K = 10$, $K = 20$ and $K = 40$, two possible cutting points, $\theta_c = 1.00$ and $\theta_c = 0.10$, and several choices of the true ability, that is, in some studies, for each replication a value of θ was drawn from a standard normal distribution, and in other studies, θ remained fixed at $\theta = -0.50$, $\theta = 0.00$, or $\theta = 0.50$.

Consider the results of Table 6. In the simulation studies giving rise to this table, the cutting score was $\theta_c = 1.00$. Notice that, in terms of mean loss, sequential testing did slightly improve upon a fixed test. In the studies to be discussed, it will become apparent that this effect increased as a function of the total number of items K; it will become apparent that for $K = 40$, this effect became quite large. Further, in Table 6 it can be seen that adaptive sequential testing does indeed improve upon sequential testing in terms of mean loss, but this effect was generally small, and it was not consistent over all three adaptive selection methods. Below, it will become apparent that the decrease of mean loss depended on the position of the cut-off score. Further, it can be seen that the decrease of mean loss was mainly due to a dramatic reduction in the proportion of testlets given. The number of correct classifications remained stable. Below, it will become apparent that the proportion of testlets given decreased further with increasing K. Finally, it can be seen that the mean loss was smallest in the cases that K testlets of one item

each were given. This result will be further corroborated in the sequel.

Table 7 contains results for a combination of a cutting score $\theta_c = 0.10$, an expected true $\theta = 0$ and a total number of items $K = 10$. This combination of a cutting score very close to the mean true ability and a small number of items produced the worst overall results. This will be further corroborated in a simulation with θ fixed. Still, a combination of K testlets and item selection at the cutting point produced the best results.

Insert Tables 8 to 11 about here

The Tables 8 to 11 contain the results for all combinations of $K = 20$ or $K = 40$ and $\theta_c = 0.10$ or $\theta_c = 1.00$. The negative relation between the number of testlets and mean loss remained apparent, and, overall, the mean loss for ASMT was slightly better. Notice that, in Table 10, the decrease of loss for the combination of 40 testlets and $\theta_c = 1.00$, displayed in the four bottom rows with respect to the loss of a fixed test, displayed in the first row, is quite dramatic, mainly due to the fact that the proportion on items given decreases to 0.10.

Insert Tables 12 to 19 about here

The Tables 12 to 19 are an overview of simulation studies of how SMT and ASMT perform for some fixed points on the ability scale. Four conditions were studied: a combination of $\theta_c = 0.10$ with $\theta = -0.50$, $\theta = 0.00$, and $\theta = 0.50$, respectively, and a combination of $\theta_c = 1.00$ and $\theta = 0.50$. Notice that distance between the true ability in the first and the latter are roughly the same; however, the reason for adding these conditions they are differently located with respect to the standard normal prior ability distribution. First consider the Tables 12 to 15, where the results for $K = 20$ are given. As above, in all tables, there is a substantial main effect on average loss of augmenting the number of testlets and a small main effect on average loss of adaptive testlet selection. Comparing the four bottom lines of the Tables 12, 13 and 14, it can be seen that a small distance between the true θ and θ_c does not necessarily produce the largest losses; however, the proportions of testlets that must be administered to attain this result are slightly larger than the proportions of testlets that must be administered in the two other cases, reported in Tables 12 and 14. Comparing Tables 12, 13 and 14 with Table 15, it can be seen that the position of θ and θ_c with respect to the prior ability distribution can have important consequences: overall, the losses dramatically decrease, and the gain from adaptive testlet selection becomes more pronounced.

The picture of the Tables 12 to 15 becomes less clear when the total number of items K is augmented to 40. In Tables 16, 17 and 18, administering 40 testlets of 1 item no longer uniformly produces the smallest loss; only Table 19 still presents the clear-cut picture of a substantial main effect for number of testlets and a small main effect for adaptive testlet selection.

Discussion

In this paper, a general theoretical framework for sequential mastery testing based on a combination of Bayesian sequential decision theory and item response theory was presented. Further, it was shown that the implication of IRT supports adaptive item and testlet selection. Then the impact of sequential testing and adaptive item selection on average loss was investigated in a number of simulation studies. It was found that sequential mastery testing does indeed lead to a considerable decrease of loss, mainly due to a significant decrease of testlets administered. The number of correct decisions remains relatively stable. The decrease of loss is positively related to the number of items in a testlet: the larger the number of testlets and the smaller the number of items in a testlet, the less the loss. The reduction of loss due to adaptive testlet selection is less pronounced. Across studies, the minimal variance criterium (14) and selection of testlets with maximum information near the cut-off point θ_c produce the best results, but the difference with the maximum information criterium is very small. Summing up, the conclusion is that the combination of Bayesian sequential decision theory and IRT framework provides a sound framework for sequential mastery testing where both the cost of test administration and the distance between the testees ability and cut-off point have to be taken into account. Finally, the merits of adaptive testing must not be exaggerated.

The general approach sketched here can be applied to several other IRT models; the main bottleneck is the computation of the expected risk defined by (6). This will present the following problems.

- For the 2-PL model, expected risk can, in principle, still be exactly computed (see, Glas and Béguin, 1996). However, this entails computation of elementary symmetric functions for every quadrature point of the grid used for evaluation of the integrals over θ. The numerical precision of this procedure is an important point of further study. Another approach might be to approximate the 2-PL model by the so-called OPLM (Verhelst and Glas, 1995). An algorithm for this approximation has been developed by Verstralen (1996). Since the OPLM is an exponential family model, expected risk can be exactly computed using elementary symmet-

ric functions. The third method is computation of expected risk via a Monte Carlo simulation, where response patterns are drawn from the posterior predictive distribution defined by (7). The number of simulated response patterns needed to obtain a reasonable approximation can be determined by comparing the results with the results of the exact method as a base line.

- For the 3-PL model, the two exact approaches for the 2-PL model are no longer feasible, and Monte Carlo simulation might be the only method for the computation of expected risk.
- Another application is the adoption of multidimensional IRT models (see, for instance, McDonald, 1997, or Reckase, 1997). Exact computation will again be confined to special cases of multidimensional models that define exponential families. Otherwise, Monte Carlo methods will be necessary.

Another point of further study is the fact that the dependence structure that might be expected when using testlets is not properly modeled yet. Above, it was assumed that local independence simultaneously holds within and over testlets, that is, all item responses are independent given θ. However, item responses within a testlet are more alike than item responses of different testlets, and it may take an hierarchical IRT model to properly describe this dependence structure. It is expected that the performance of sequential testing might suffer from these additional sources of variation, but not conclusive assertions can be made until further research is done.

References

DeGroot, M.H. (1970). *Optimal statistical decisions.* New York: McGraw-Hill.

De Gruijter, D.N.M., & Hambleton, R.K. (1984). On problems encountered using decision theory to set cutoff scores. *Applied Psychological Measurement, 8,* 1-8.

Glas, C.A.W. & Béguin, A.A. (1996). *Appropriateness of IRT observed score equating.* Research Report 96-04, University of Twente, Enschede.

Kingsbury, G.G., & Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D.J. Weiss (Ed.): *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257-283). New York: Academic Press.

Lehman, E.L. (1986). *Testing statistical hypothesis. (second edition).* New York: Wiley.

, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367-386.

McDonald, R.P. (1997). Normal-ogive multidimensional model. In: W.J. van der Linden and R.K. Hambleton (eds.). *Handbook of Modern Item Response Theory.* (pp.257-269). New York: Springer.

van der Linden, W.J. (1990). Applications of decision theory to test-based decision making. In R.K. Hambleton & J.N. Zaal (Eds.): *New developments in testing: Theory and applications,* 129-155. Boston: Kluwer.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In: W.J. van der Linden and R.K. Hambleton (eds.). *Handbook of Modern Item Response Theory.* (pp.271-286). New York: Springer.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with non-equivalent testlets. *Applied Psychological Measurement, 16,* 65-76.

Spray, J.A., & Reckase, M.D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21,* 405-414.

Verhelst, N.D., & Glas, C.A.W. (1995). The generalized one parameter model: OPLM. In: G.H. Fischer & I.W. Molenaar (eds.). *Rasch models: their foundations, recent developments and applications* . New York: Springer.

Verhelst, N.D., Glas, C.A.W. & van der Sluis, A. (1984). Estimation problems in the Rasch model: the basic symmetric functions. *Computational Statistics Quarterly, 1,* 245-262.

Verstralen, H.H.F.M. (1996). Optimal Integer Category Weights in the OPLM and

BEST COPY AVAILABLE

GPCM. Measurement and Research Department Reports, 95-2. Cito: Arnhem.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Table 1
Standard Errors for θ

number of items	p-value 0.50 SE	0.25 SE	0.10 SE
10	.6325	.7303	1.0541
20	.4472	.5164	.7454
30	.3651	.4216	.6086
40	.3162	.3651	.5270
50	.2828	.3266	.4714
60	.2582	.2981	.4303
70	.2390	.2760	.3984
80	.2236	.2582	.3727
90	.2108	.2434	.3514
100	.2000	.2309	.3333

Table 2
Gain in Proportion of Correct Decisions
K=10

	Proportion Incorrect		
θ_c	p=0.50	p=0.25	Gain
.0	.5000	.5000	.0000
.1	.4372	.4455	.0084
.2	.3759	.3921	.0162
.3	.3176	.3406	.0230
.4	.2635	.2919	.0284
.5	.2146	.2468	.0322
.6	.1714	.2057	.0343
.7	.1342	.1689	.0347
.8	.1030	.1367	.0337
.9	.0774	.1089	.0315
1.0	.0569	.0855	.0285
1.1	.0410	.0660	.0250
1.2	.0289	.0502	.0213
1.3	.0199	.0375	.0176
1.4	.0134	.0276	.0142
1.5	.0089	.0200	.0111
1.6	.0057	.0142	.0085
1.7	.0036	.0100	.0064
1.8	.0022	.0069	.0046
1.9	.0013	.0046	.0033

Table 3
Gain in Proportion of Correct Decisions
K=20

	Proportion Incorrect		
θ_c	p=0.50	p=0.25	Gain
.0	.5000	.5000	.0000
.1	.4115	.4232	.0117
.2	.3274	.3493	.0219
.3	.2512	.2806	.0295
.4	.1855	.2193	.0337
.5	.1318	.1665	.0347
.6	.0899	.1226	.0328
.7	.0588	.0876	.0289
.8	.0368	.0607	.0238
.9	.0221	.0407	.0186
1.0	.0127	.0264	.0137
1.1	.0070	.0166	.0096
1.2	.0036	.0101	.0064
1.3	.0018	.0059	.0041
1.4	.0009	.0034	.0025
1.5	.0004	.0018	.0014
1.6	.0002	.0010	.0008
1.7	.0001	.0005	.0004
1.8	.0000	.0002	.0002
1.9	.0000	.0001	.0001

Table 4
Gain in Proportion of Correct Decisions
K=40

	Proportion Incorrect		
θ_c	p=0.50	p=0.25	Gain
.0000	.5000	.5000	.0000
.1000	.3759	.3921	.0162
.2000	.2635	.2919	.0284
.3000	.1714	.2057	.0343
.4000	.1030	.1367	.0337
.5000	.0569	.0855	.0285
.6000	.0289	.0502	.0213
.7000	.0134	.0276	.0142
.8000	.0057	.0142	.0085
.9000	.0022	.0069	.0046
1.0000	.0008	.0031	.0023
1.1000	.0003	.0013	.0010
1.2000	.0001	.0005	.0004
1.3000	.0000	.0002	.0002
1.4000	.0000	.0001	.0001
1.5000	.0000	.0000	.0000
1.6000	.0000	.0000	.0000
1.7000	.0000	.0000	.0000
1.8000	.0000	.0000	.0000
1.9000	.0000	.0000	.0000

Table 5
Correct Decisions and Item Selection Method
$\theta_c = 1.00$, 5000 replications

	selection	K=10		K=20		K=40	
ability	method	WML	EAP	WML	EAP	WML	EAP
$\theta = 0.50$	random	.77	.90	.84	.93	.92	.95
	$\beta = \theta_c$	.86	.96	.91	.96	.96	.98
	$\beta = \theta$	.83	.95	.87	.94	.96	.96
	$\beta = \hat{\theta}_s$	.80	.79	.86	.83	.94	.91
$\theta \sim N(0, 1)$	random	.81	.89	.82	.87	.82	.84
	$\beta = \theta_c$	.85	.91	.85	.88	.84	.86
	$\beta = \theta$	.81	.88	.82	.86	.83	.84
	$\beta = \hat{\theta}_s$	.81	.81	.82	.81	.83	.83

Table 6
Relation Between Selection Method and Loss
K=10, $\theta \sim N(0, 1)$, $\theta_c = 1.00$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	10	fixed test	.90	1.00	.1546
2	5	sequential	.90	.76	.1417
2	5	max info	.90	.76	.1242
2	5	min variance	.91	.74	.1217
2	5	cutting point	.89	.75	.1297
10	1	sequential	.89	.46	.1091
10	1	max info	.87	.42	.1219
10	1	min variance	.91	.41	.0920
10	1	cutting point	.87	.43	.1137

Table 7
Relation between Selection Method and Loss
K=10, $\theta \sim N(0,1)$, $\theta_c = 0.10$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	10	fixed test	.84	1.00	.1819
2	5	sequential	.81	.91	.1855
2	5	max info	.83	.91	.1837
2	5	min variance	.79	.91	.2109
2	5	cutting point	.82	.89	.1795
10	1	sequential	.80	.86	.2037
10	1	max info	.81	.83	.1900
10	1	min variance	.81	.81	.1910
10	1	cutting point	.83	.85	.1657

Table 8
Relation between Selection Method and Loss
K=20, $\theta \sim N(0, 1)$, $\theta_c = 1.00$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	20	fixed test	.90	1.00	.2440
2	10	sequential	.91	.63	.1645
2	10	max info	.91	.64	.1683
2	10	min variance	.92	.63	.1589
2	10	cutting point	.93	.64	.1554
4	5	sequential	.89	.42	.1373
4	5	max info	.91	.41	.1209
4	5	min variance	.91	.42	.1255
4	5	cutting point	.91	.41	.1245
20	1	sequential	.89	.26	.1119
20	1	max info	.91	.25	.0979
20	1	min variance	.92	.27	.0957
20	1	cutting point	.90	.27	.0968

Table 9
Relation between Selection Method and Loss
K=20, $\theta \sim N(0,1)$, $\theta_c = 0.10$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	20	fixed test	.88	1.00	.2506
2	10	sequential	.86	.70	.2030
2	10	max info	.87	.71	.1958
2	10	min variance	.85	.70	.2050
2	10	cutting point	.84	.70	.2060
4	5	sequential	.84	.55	.1824
4	5	max info	.86	.60	.1816
4	5	min variance	.86	.58	.1735
4	5	cutting point	.85	.55	.1769
20	1	sequential	.85	.54	.1700
20	1	max info	.83	.49	.1668
20	1	min variance	.87	.49	.1492
20	1	cutting point	.85	.46	.1625

Table 10
Relation between Selection Method and Loss
K=40, $\theta \sim N(0,1)$, $\theta_c = 1.00$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	40	fixed test	.92	1.00	.4316
4	10	sequential	.90	.30	.1577
4	10	max info	.91	.28	.1515
4	10	min variance	.92	.29	.1493
4	10	cutting point	.91	.28	.1661
10	4	sequential	.90	.19	.1148
10	4	max info	.90	.18	.1107
10	4	min variance	.92	.17	.1002
10	4	cutting point	.91	.19	.1169
40	1	sequential	.90	.10	.1012
40	1	max info	.89	.11	.0949
40	1	min variance	.91	.10	.0883
40	1	cutting point	.89	.10	.1023

Table 11
Relation between Selection Method and Loss
K=40, $\theta \sim N(0,1)$, $\theta_c = 0.10$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	40	fixed test	.91	1.00	.4263
4	10	sequential	.84	.30	.1972
4	10	max info	.87	.35	.1846
4	10	min variance	.88	.33	.1836
4	10	cutting point	.87	.35	.1884
10	4	sequential	.82	.20	.1678
10	4	max info	.82	.20	.1708
10	4	min variance	.85	.19	.1470
10	4	cutting point	.82	.21	.1742
40	1	sequential	.85	.20	.1540
40	1	max info	.85	.19	.1454
40	1	min variance	.84	.19	.1603
40	1	cutting point	.86	.20	.1420

Table 12
Relation between Selection Method and Loss
K=20, $\theta = -0.50$, $\theta_c = 0.10$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	20	sequential	.91	1.00	.2846
2	10	sequential	.91	.77	.2389
2	10	max info	.89	.77	.2531
2	10	min variance	.90	.75	.2396
2	10	cutting point	.90	.75	.2442
4	5	sequential	.87	.60	.2370
4	5	max info	.90	.61	.2118
4	5	min variance	.91	.59	.2019
4	5	cutting point	.89	.61	.2212
20	1	sequential	.88	.56	.2187
20	1	max info	.90	.55	.2035
20	1	min variance	.89	.54	.2093
20	1	cutting point	.91	.50	.1822

Table 13
Relation between Selection Method and Loss
K=20, $\theta = 0.00$, $\theta_c = 0.10$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	20	sequential	.59	1.00	.2610
2	10	sequential	.58	.82	.2279
2	10	max info	.57	.83	.2289
2	10	min variance	.56	.83	.2324
2	10	cutting point	.59	.84	.2286
4	5	sequential	.56	.71	.2077
4	5	max info	.60	.78	.2153
4	5	min variance	.63	.72	.2004
4	5	cutting point	.60	.71	.2026
20	1	sequential	.62	.68	.1931
20	1	max info	.57	.65	.1960
20	1	min variance	.59	.67	.1950
20	1	cutting point	.60	.63	.1849

Table 14
Relation between Selection Method and Loss
K=20, $\theta = 0.50$, $\theta_c = 0.10$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	20	sequential	.82	1.00	.3092
2	10	sequential	.81	.75	.2607
2	10	max info	.80	.76	.2708
2	10	min variance	.84	.76	.2500
2	10	cutting point	.82	.77	.2604
4	5	sequential	.78	.68	.2693
4	5	max info	.75	.73	.2936
4	5	min variance	.76	.73	.2888
4	5	cutting point	.80	.70	.2624
20	1	sequential	.75	.64	.2762
20	1	max info	.78	.62	.2544
20	1	min variance	.79	.62	.2482
20	1	cutting point	.77	.59	.2574

Table 15
Relation between Selection Method and Loss
K=20, $\theta = 0.50$, $\theta_c = 1.00$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	20	fixed test	.92	1.00	.2577
2	10	sequential	.94	.71	.1905
2	10	max info	.96	.72	.1784
2	10	min variance	.95	.71	.1764
2	10	cutting point	.92	.73	.2091
4	5	sequential	.93	.49	.1493
4	5	max info	.96	.46	.1248
4	5	min variance	.95	.47	.1309
4	5	cutting point	.93	.48	.1525
20	1	sequential	.92	.36	.1297
20	1	max info	.95	.32	.1010
20	1	min variance	.96	.34	.0973
20	1	cutting point	.94	.36	.1161

Table 16
Relation between Selection Method and Loss
K=40, $\theta = -0.50$, $\theta_c = 0.10$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	40	sequential	.98	1.00	.4216
4	10	sequential	.85	.32	.2661
4	10	max info	.93	.35	.2028
4	10	min variance	.92	.35	.2153
4	10	cutting point	.93	.36	.2070
10	4	sequential	.90	.20	.1685
10	4	max info	.91	.20	.1636
10	4	min variance	.89	.19	.1798
10	4	cutting point	.90	.20	.1687
40	1	sequential	.89	.22	.1835
40	1	max info	.88	.20	.1873
40	1	min variance	.88	.20	.1873
40	1	cutting point	.88	.21	.1913

Table 17
Relation between Selection Method and Loss
K=40, $\theta = 0.00$, $\theta_c = 0.10$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	40	sequential	.67	1.00	.4495
4	10	sequential	.53	.35	.2092
4	10	max info	.61	.42	.2259
4	10	min variance	.62	.44	.2319
4	10	cutting point	.61	.44	.2351
10	4	sequential	.63	.27	.1630
10	4	max info	.63	.24	.1511
10	4	min variance	.61	.26	.1606
10	4	cutting point	.63	.27	.1637
40	1	sequential	.60	.28	.1708
40	1	max info	.60	.24	.1578
40	1	min variance	.60	.24	.1578
40	1	cutting point	.60	.27	.1685

Table 18
Relation between Selection Method and Loss
K=40, $\theta = 0.50$, $\theta_c = 0.10$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	40	sequential	.85	1.00	.4876
4	10	sequential	.80	.33	.2522
4	10	max info	.79	.40	.2868
4	10	min variance	.78	.40	.2879
4	10	cutting point	.82	.42	.2719
10	4	sequential	.72	.26	.2751
10	4	max info	.72	.24	.2640
10	4	min variance	.69	.23	.2817
10	4	cutting point	.71	.26	.2790
40	1	sequential	.77	.25	.2416
40	1	max info	.75	.23	.2433
40	1	min variance	.75	.23	.2433
40	1	cutting point	.74	.26	.2594

Table 19

Relation between Selection Method and Loss

K=40, $\theta = 0.50$, $\theta_c = 1.00$, 1000 replications

number of testlets	items per testlet	selection method	proportion correct decisions	proportion testlets given	mean loss
1	40	fixed test	.97	1.00	.4225
4	10	sequential	.92	.31	.1846
4	10	max info	.93	.31	.1773
4	10	min variance	.92	.29	.1732
4	10	cutting point	.91	.31	.1892
10	4	sequential	.93	.22	.1409
10	4	max info	.96	.19	.1071
10	4	min variance	.96	.19	.1113
10	4	cutting point	.94	.23	.1370
40	1	sequential	.94	.13	.0962
40	1	max info	.96	.12	.0778
40	1	min variance	.95	.13	.0868
40	1	cutting point	.96	.11	.0737

Author Note

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this paper are those of the author and do not necessarily reflect the position or policy of LSAC.

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

RR-98-09 B.P. Veldkamp, *Multiple Objective Test Assembly Problems*

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

RR-98-07 W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

43